Stat Comput (2018) 28:201–217
https://doi.org/10.1007/s11222-017-9726-x

# Automatic specification of piecewise linear additive models: application to forecasting natural gas demand

**Alberto Gascón**[1] · **Eugenio F. Sánchez-Úbeda**[1]

Received: 22 July 2016 / Accepted: 5 January 2017 / Published online: 25 January 2017


**Abstract** When facing any forecasting problem not only is accuracy on the predictions sought. Also, useful information about the underlying physics of the process and about the relevance of the forecasting variables is very much appreciated. In this paper, it is presented an automatic specification procedure for models that are based on additivity assumptions and piecewise linear regression. This procedure allows the analyst to gain insight about the problem by examining the automatically selected model, thus easily checking the validity of the forecast. Monte Carlo simulations have been run to ensure that the model selection procedure behaves correctly under weakly dependent data. Moreover, comparison over other well-known methodologies has been done to evaluate its accuracy performance, both in simulated data and in the context of short-term natural gas demand forecasting. Empirical results show that the accuracy of the proposed model is competitive against more complex methods such as neural networks.

**Keywords** Generalized additive models · Prediction · Natural gas demand · Short-term forecasting · Piecewise linear models · Nonlinear modeling

✉ Eugenio F. Sánchez-Úbeda
eugenio.sanchez@iit.comillas.edu

Alberto Gascón
alberto.gascon@iit.comillas.edu

[1] Institute for Research in Technology (IIT), Technical School of Engineering (ICAI), Comillas Pontifical University, Santa Cruz de Marcenado, 26, 28015 Madrid, Spain

## 1 Introduction

Nonlinear time series modeling experimented a tipping point in attention two decades ago. Motivated by the increasing awareness of the limitations of linear models (see Tong 1993 for instance) and the exponential raise of computational power available, varied new techniques arose with the aim of solving the handicaps of linear approaches. Despite this effort, nonlinear methodologies have not penetrated into the industry as a clear dominant approach over linear methods. In 2004, Clements et al. (2004) pointed out that nonlinear approaches "… were not mimicking reality any better than simpler linear approximations". Not so much seems to have changed in the last decade, and very often the industry struggles to handle nonlinearities on a clear manner. Just for providing an example, more recently in 2011 Cruz et al. (2011) found in periodic dynamic regression model the best approximation for electricity prices forecasting, beating classical methods such as feed-forward neural networks. We still see some industrial applications of linear models for time series that exhibit clear nonlinearities and this election may probably be due to the robustness and simplicity of linear models and more specially to its interpretability.

For example, in the context of energy forecasting, the relationship between consumption and outdoor air temperature needs to be considered due to its relevance. This complex relationship depends on the type of consumed energy (electricity or natural gas) and on the climate characteristics of the geographical area considered. Basically, the electric consumption increases both for decreasing and increasing temperatures, being the response asymmetric and clearly nonlinear due to the use of heating appliances in winter and air conditioners in summer (for details see Muñoz et al. 2010). Natural gas consumption has a similar behavior, but it does not increase for increasing temperatures (see, e.g., Szoplik

2015). Furthermore, there is a dynamic effect due to the thermal inertia of buildings, as well as saturation effects because of the limited capacity of the installed heating and cooling appliances. This nonlinear relationship is typically modeled in the literature by using as input variables, instead of temperature, several degree-days transformations of it, such as the heating degree-days HDD $= \max(R^H - T, 0)$ and the cooling degree-days CDD $= \max(T - R^C, 0)$, defined as the difference between the actual mean temperature $T$ and a reference temperature ($R^H$ and $R^C$ for winter and summer, respectively). Note that the reference temperatures should be adequately selected in order to separate correctly the cold and heat branches of the consumption–temperature relationship. The simplest approach uses a fixed reference value for both HDD and CDD (e.g., 65 °F in the USA, 15 °C in Spain).

In a world where we are tending to have more and more available data, extracting useful information with high interpretable models allows us to make understandable decisions, which helps to justify these decisions when necessary and therefore are more likely to be used in the industry under uncertainty. Also, the capability of automatically identifying the correct structure of a model is an important practical feature of any approach, in particular, selecting the important lags of a univariate model or the relevant exogenous variables of a multivariate model is a desirable aspect and this way, great amounts of data are more easily tractable. In this paper, an automatic specification procedure (the learning algorithm) is presented for a model based on global linear-splines regression. Following inspiring classic names in the context of generalized additive models such as Following inspiring classic names in the context of generalized additive models such as MARS (Multivariate Adaptive Regression Splines), or SMART (Smooth Multiple Additive Regression Technique), we will call this approach SNAKE (Spline Nonlinear Additive with varying Knots Estimation). As we will see, the benefit of this approach is not only the automatic specification of the model, but also due to the simplicity of the structure of the model itself, the ease of interpreting the results. This procedure is applicable to virtually any kind of time series problem, while we have studied the problem of short-term forecasting natural gas demand. In the context of energy demand forecasting, additive models have proven to be a reasonable approximation, yielding accurate results (Fan and Hyndman 2012 for instance used this kind of approach for short-term electricity load). Although electricity demand has traditionally received much more attention than natural gas, in recent years, it is possible to find an increasing number of publications on the topic. For example, Soldo (2012) can be revised for an analysis on natural gas demand approaches. More recently, Zhu et al. (2015) and Szoplik (2015) use neural networks and support vector regression, respectively.

When automatically modeling a nonlinear time series, such as those usually found in the context of natural gas demand, two essential aspects have to be considered. First, some kind of structure has to be chosen. If too rigid, the model may not be able to capture all the essential particularities of the data, but if a very flexible structure is chosen, care must be taken to avoid overfitting. Second, an appropriate method must be applied for selecting the correct dimension of the model (i.e., choosing the significant lag values and exogenous variables). For these reasons, additive models are a common compromise solution between general nonparametric approaches and simple linear models.

Nonparametric approaches with automatic lag selection were addressed by Cheng and Tong (1992), Vieu (1995), Yao and Tong (1994) with a cross-validation approach, whereas Tjøstheim and Auestad (1994) used a nonparametric version of the final prediction error (FPE) criterion of Akaike (see Akaike 1970), all based on the Nadaraya–Watson estimator. However, all these lag selection methods are computationally intensive owing to the local nature of kernel smoothing. Moreover, they have found to behave poorly on low sample size situations (see Guo and Shintani 2011; Tschernig and Yang 2000), and they may suffer from some overfitting bias in these conditions. For this reason, Tschernig and Yang (2000) introduced a correction on the FPE criterion with local estimators and Guo and Shintani (2011) pointed out that, even with this correction, when the sample is too small, there was still a noteworthy bias which could be diminished by introducing additivity in the autoregressive function.

On the other hand, if the errors are supposed to follow some a-priori known distribution, parametric models can be applied and estimated through maximum likelihood. Chen and Tsay (1993) proposed using the alternating conditional expectation (ACE) and BRUTO algorithms of Hastie et al. (2009) to identify the nonlinear functions and to perform variable selection for additive autoregression. Chen and Tsay pointed out that the convergence of the algorithm could be slow when dealing with highly correlated observations, and Huang and Yang (2004) argued that the method lacks a theoretical justification, while no systematic simulation study has been done to evaluate its performance.

Also, Lewis and Stevens (1991) proposed a forecasting model based on the regression methodology of MARS (see Hastie et al. 2009). MARS model is characterized by permitting nonadditive interaction between variables if necessary, but it is not restricted to enforce this interaction. Although MARS method can perform automatic variable selection, its ability to identify the set of significant variables (or lags) is not clear. All these deficiencies were pointed out by Huang and Yang (2004) who developed a deep theoretical study on using BIC criterion for lag selection and proved their theoretical results with Monte Carlo simulations for varied time series. The approach presented in Huang and Yang (2004) is based on polynomial spline global regression and used the BIC for lag selection. A very neat demonstration is pre-

sented showing that under appropriate assumptions, the BIC approach is a consistent selection rule. Basically, this conclusion lies on general assumptions on the data-generating process (mainly strongly mixing, $\alpha$-mixing, conditions) and specifically implies that it is necessary to let the number of knots (i.e., the number of spline basis functions for each variable) increase with the sample size. This implication, although not being specially misbegotten, is fairly questionable. One could expect that the data would define the complexity (number of knots) of the model, regardless of the sample size. A clear example is linear autoregressive time series, where the optimal dimension of the space of spline functions is minimum and particularly independent of the sample size. Moreover, in Huang and Yang (2004) the knots are assumed to be equally placed among the data, and therefore, it is expected that minor changes in the knot position have little effect on the model performance. If this condition does not hold (e.g., when threshold values of the data represent a significant change in the time series behavior), a more intelligent algorithm should be employed for spline regression.

In this paper, the SNAKE approach is proposed, based on global linear-splines regression, similar to Huang and Yang (2004), but that automatically selects the adequate number and position of the knots as well as the significant lags and variables to be included in the model. The SNAKE model is a time series adaptation of the ORTHO model, which was originally proposed in 1999 as a novel supervised automatic learning model based on the ideas of projection pursuit and linear splines (Sánchez-Úbeda 1999; Sánchez-Úbeda and Wehenkel 2000) and was successfully applied in the context of pure regression in power system security assessment (see Wehenkel 1998). Basically, the ORTHO model offers interpretability, capability to identify the input variables that influence most strongly the output, and modeling flexibility. In particular, this model belongs to the generalized additive model class (see Hastie et al. 2009), where the output is estimated by a weighted sum of terms representing the contribution of each input variable separately. Note that this simplification of the functional form means gaining interpretability of the model, but it has a cost since not any underlying function can be represented in this way (i.e., these models are not universal approximators).

In the following sections, we describe the structure of the nonlinear SNAKE model (Sect. 2) and the automatic learning algorithm proposed to fully identify its structure as well as to estimate all its parameters (Sect. 3). Next, in order to ascertain the goodness of the SNAKE model, we perform in Sect. 5 a Monte Carlo simulation study using different synthetic time series to empirically compare the proposed approach with other well-known benchmark methodologies. Finally, before concluding, in Sect. 6 we compare the results obtained when forecasting day ahead natural gas consumption in Spain.

## 2 Model structure

Let us begin with a formal statement of the forecasting problem. In time series modeling, a regression function is a general relationship between some response variable $y_t$ and a vector of $d$ explanatory variables $\boldsymbol{w_t} = (w_t^1, \ldots, w_t^d)$ which can be either lagged values of $y_t$, errors, any exogenous variable, or lagged exogenous variables. The response is assumed to depend on the inputs through the relationship

$$y_t = \phi(\boldsymbol{w_t}) + \varepsilon_t, \tag{1}$$

where $\phi(\cdot)$ shows the expected value of $y_t$ as a function of $\boldsymbol{w_t}$. The deviations of $y_t$ around its expectation value, caused by non-observed input variables and other random effects are represented by a noise component $\varepsilon_t$, typically assumed to be uncorrelated, normally distributed with zero mean and unknown variance.

To deal with this regression problem, the SNAKE model can be stated as

$$S_s = \beta_0 + \sum_{i \in s} \beta_i \mathrm{LHM}_i(w_t^i), \tag{2}$$

where $s$ is the subset of variables $s \subset \{1, \ldots, d\}$ considered mathematically relevant for the given problem, $\mathrm{LHM}_i(\cdot)$ are one-dimensional linear hinges models, explained below, and the $\beta_i$ parameters reflect the contribution of each term. This way, each term in Eq. (2) appertains to a one-dimensional model selecting a particular input variable. For the sake of interpretability, the parameters $\boldsymbol{\beta}$ are readjusted to agree with the constraints $\{mean(\mathrm{LHM}_i(\cdot)) = 0 \;\; \& \;\; std(\mathrm{LHM}_i(\cdot)) = 1\}$. This allows a simpler interpretation of the final model.

The problem of variable selection arises when determining the correct subset $s$ such that all the significant variables are included but no other variable is. Correct identification of $s$ will be addressed in Sect. 3.2 and on the simulation studies.

### The Linear Hinges Model structure

The LHM is a one-dimensional piecewise linear model completely defined by a set of $K$ internal knots or *hinges* $(k_j, h_j)$, plus two external nodes that specify the boundary conditions (the extrapolation slopes) as shown in Fig. 1. The mathematical expression of the model is

$$\mathrm{LHM}(x) = \begin{cases} h_0 + m_1(x - k_0) & x < k_1 \\ h_{j-1} + m_j(x - k_{j-1}) & k_{j-1} \le x < k_j \\ h_K + m_{K+1}(x - k_K) & x \ge k_K \end{cases} \tag{3}$$

where $j = 2, \ldots, K$ and $m_i = (h_i - h_{i-1})/(k_i - k_{i-1})$ is the slope of the linear piece between knots $(k_{i-1}, h_{i-1})$ and $(k_i, h_i)$, with $i = 1, \ldots, K+1$.

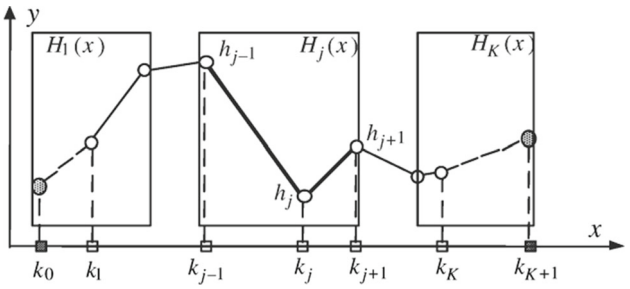


**Fig. 1** Linear hinges model (LHM), a piecewise linear model consisting of $K$ internal knots given by coordinates $(k_j, h_j)$

The LHM (Sánchez-Úbeda and Wehenkel 1998; Sánchez-Úbeda 1999) is the basic brick of the SNAKE model, its elementary piece (see Eq. 2), the same way as multilayer perceptrons consist of hyperbolic tangent functions as the basic building blocks. In fact, the SNAKE model can be thought as a natural multidimensional extension of the LHM.

Although this kind of models are well known in the literature, the main advantage of the LHM compared to other similar ones is that its learning algorithm automatically selects the number and location of the knots, adapting its complexity to the quality and availability of the data and thus allowing to describe a wide range of functional forms. This feature makes the LHM very flexible, able to produce adequate models in many different situations.

## 3 Learning algorithm

The goal of the learning algorithm is to estimate a function $g(\cdot)$ that performs a good approximation of the true underlying function $\phi(\cdot)$ of Eq. (1). In order to do this, some criterion must be defined to say that $g(\cdot)$ is close enough to $\phi(\cdot)$. A very common criterion and the one that will be used here is to minimize the mean squared error (MSE) in the learning time period. We can define the MSE as

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}[y_t - g(\boldsymbol{w_t})]^2, \tag{4}$$

where $n$ is the length of the time period used for learning. In this section, first the learning algorithm for the LHM will be explained. Later, the SNAKE model can be easily constructed by orthogonal combination of LHMs.

### 3.1 LHM's learning algorithm

Exact details of the adjustment of the LHM can be found at Sánchez-Úbeda and Wehenkel (1998), Sánchez-Úbeda (1999), Sánchez-Úbeda and Berzosa (2007), but briefly speaking, the learning algorithm consists of two steps. In the first step, an initial model is built using the noise-filtered data obtained with the supersmoother, a scatterplot smoothing algorithm proposed by Friedman (1984) and based on local linear averaging with variable span. After smoothing the data, knots are added on a growing sequence until a maximum complexity is reached. Then a backward pruning sequence is done. Model selection is done through evaluating the model on a so-called pruning set; a preselected subset of the data not used for estimating the model parameters but for identifying its correct complexity. Finally, in the second step, once that the model complexity has been chosen and that some initial $x$-positions are suggested for the knots, the LHM is refitted by using the whole set of the original scatterplot data.

Note that other learning algorithms have been proposed for piecewise linear fitting. Genetic algorithms in particular have proven to be specially useful for this kind of problems (see Gascón et al. 2011), but the increase in accuracy is not paid off by the computational cost of these techniques, specially when SNAKE's models are expected to adjust many LHMs for a typical problem.

### 3.2 SNAKE's learning algorithm

The learning strategy of the SNAKE model automatically selects the most relevant variables, that is, those that have a greater influence in the response. In fact, this automatic selection of variables is one of the main features of the proposed approach. The core of the learning strategy is the minimization of the MSE using a particular implementation of the backfitting algorithm (see Sánchez-Úbeda 1999). Basically, according to Eq. (2), two main groups of parameters have to be estimated: the subset $s$ of variables included, and the form of the one-dimensional model for each variable. To carry out this estimation, the learning algorithm has three main stages: forward stage, backward stage, and variable selection.

*Forward stage* First, variables are included in the model on a forward stepwise approach. Beginning with $\beta_0$, candidate variables are sequentially selected to be included in the model. Thus, at each step, a set of auxiliary models are built by incorporating to the previous model each remaining candidate variable, and the auxiliary model with lower MSE is selected for the next step. This growing process stops when either all possible variables or a parametrized maximum number of them have been included.

Each univariate term of the model is adjusted removing the effects of all the other variables from $y_t$ and thus obtaining the partial residuals. Consequently, for any candidate variable $w_t^c$ we have

$$\beta_c\,\text{LHM}_c(w_t^c) \approx E\left[y_t - \beta_0 - \sum_{i\in s}\beta_i\,\text{LHM}_i\,(w_t^i)\;\middle|\; w_t^c\right] \tag{5}$$

Once that the candidate variable has been included in the model, the unexplained noise observed from the other orthogonal projections may have changed in level but also in shape (see Sánchez-Úbeda 1999). For this reason, according to the backfitting strategy, it is recommended to refit all the previous LHMs. The main objective of this refitting is to adjust the $x$-positions of the knots and the number of them for each projection. The height of each knot is calculated on a final step by optimizing the global surface. Despite the iterative nature of this procedure, computational efficiency is one of the main advantages of the SNAKE, since the optimization process required at each step is very fast.

*Backward stage* Second, starting with the last model of the previous stage, importances of each variable are calculated in order to evaluate its significance. These importances give a measure of $\partial y/\partial w_i$ and have been taken from Sánchez-Úbeda (1999) but with a weighting correction that takes into account the number of points $n_j$ that satisfy $k_{j-1} \leq w_t^i < k_j$ (i.e., the different linear pieces). This importance is calculated as

$$I_i = \frac{\sigma_i \beta_i}{n} \sum_{j=1}^{K_i} n_j \left| m_j \right|, \tag{6}$$

where $j$ ranges for the different linear pieces in LHM$_i$, $\sigma_i$ is the variance of $w_t^i$, and $m_j$ is the slope of the linear piece ending in $(k_j, h_j)$.

The terms with lower importance are removed first from the model, on a sequential pruning process. Once that all variables have been removed, a sequence of models is provided to the next decision stage, who must select one of these models.

*Variable selection* The original ORTHO learning algorithm shares many characteristics with the context of decision trees. In particular, in decision trees there is a need to decide the extension of the tree, such that it has as many final leafs as required but keeping robustness in mind. In fact, each LHM can be easily seen as a one-dimensional regression tree, as shown in Fig. 2.

For selecting the correct complexity of decision trees, Breiman et al. (1984) introduced the *'1 Standard Error Rule'* (1-SER), based on a cross-validation approach. For this reason, the whole learning set must be divided into, at least, two different sets, the growing set (GS) and the pruning set (PS). The GS is used to grow a large tree, whereas the PS is used during the pruning stage to select the final complexity. Both the LHM and the ORTHO model use this kind of approach to select their complexity (the LHM chooses the number of knots, whereas the ORTHO model chooses the number of terms in the model). In the SNAKE approach,

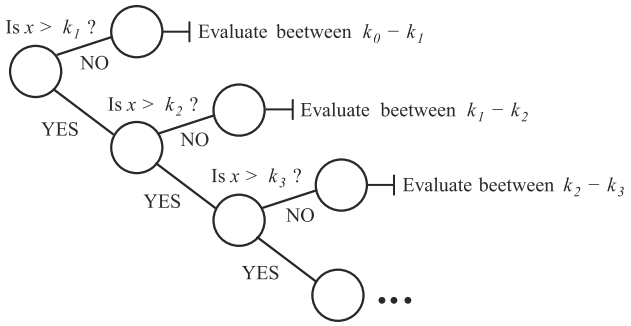


**Fig. 2** Linear hinges model seen as a regression tree where the splits of the internal nodes are given by the $x$-coordinates of the knots, and the terminal nodes (*leaves*) consist of a particular straight line segment

however, we will see that other selection method may produce better results.

The 1-SER method implies that each pruned model is evaluated on an independent sample, the PS, which has not been seen by the model while growing, and the generalization MSE for each pruned model is estimated by

$$\mathrm{MSE}(S_s, \mathrm{PS}) = \frac{1}{N_{\mathrm{PS}}} \sum_{\forall (\boldsymbol{w_t}, y_t) \in \mathrm{PS}} (y_t - S_s(\boldsymbol{w_t}))^2, \tag{7}$$

where $S_s$ is the SNAKE model, $N_{\mathrm{PS}}$ is the number of samples in the PS, and $(\boldsymbol{w_t}, y_t)$ are the data in the PS. Then, the best pruned model will be the one minimizing this MSE. The 1-SER deals with the errors that can appear in the estimation of the MSE given by Eq. (7) because of using a finite pruning set. This rule allows choosing the simplest model whose accuracy is comparable to the minimal estimated MSE, taking into account the uncertainties of the estimate MSE$(S_s, \mathrm{PS})$. These uncertainties can be considered estimating its standard error. An expression for this standard deviation can be derived (see Breiman et al. 1984) using standard statistics as

$$\sigma^2_{\mathrm{MSE}}(S_s, \mathrm{PS}) = -\frac{1}{N_{\mathrm{PS}}} \mathrm{MSE}^2(S_s, \mathrm{PS}) + \frac{1}{N^2_{\mathrm{PS}}} \sum_{\forall (\boldsymbol{w_t}, y_t) \in \mathrm{PS}} (y_t - S_s(\boldsymbol{w_t}))^4 \tag{8}$$

Then, using this rule the best pruned model $S_s^*$ is selected as the simplest model for which the MSE$(S_s^*, \mathrm{PS})$ is not larger than the minimal MSE$(S_s^{\min}, \mathrm{PS})$ plus its standard deviation. In other words, the $S_s^*$ is the simplest model satisfying

$$\mathrm{MSE}(S_s^*, \mathrm{PS}) \leq \mathrm{MSE}(S_s^{\min}, \mathrm{PS}) + \sigma_{\mathrm{MSE}}(S_s^{\min}, \mathrm{PS}), \tag{9}$$

being $S_s^{\min}$ the pruned SNAKE model for which the estimated MSE is the minimal.

The 1-SER is not as common in the context of time series as other criteria such as the Akaike information criterion

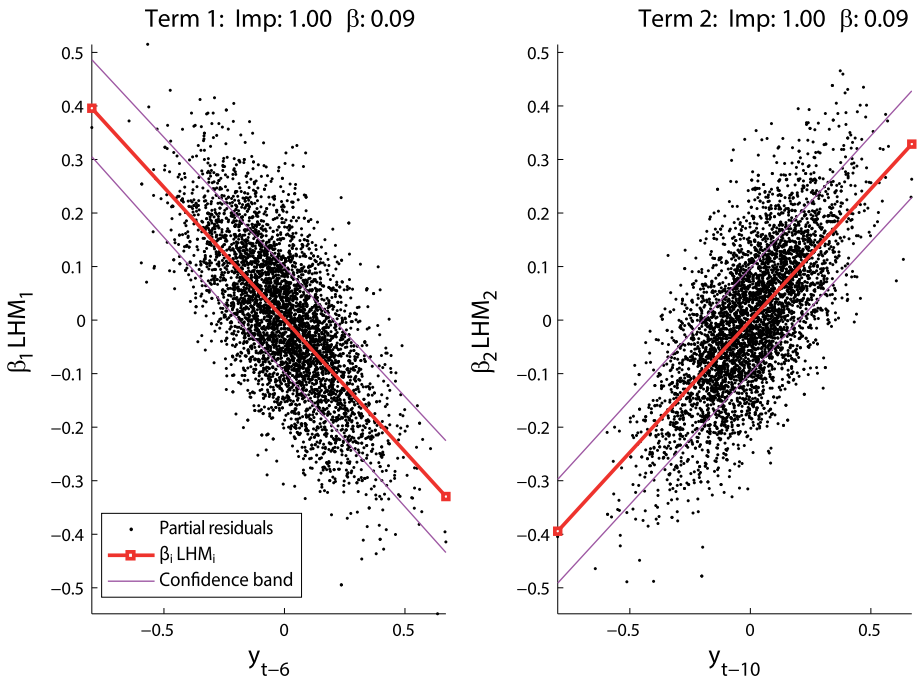


**Fig. 3** SNAKE model of a linear process. This model has no constant term ($\beta_0 = 0$), consisting of two equally important linear terms: $y_t \approx 0.09\text{LHM}_1(y_{t-6}) + 0.09\text{LHM}_2(y_{t-10})$. Note that these input variables have been automatically chosen from the first 10 lags

(AIC) or the Bayes information criterion (BIC), also known as Schwarz criterion. Both of these criteria are based on MSE minimization, but penalizing the number of parameters in the model. It is well known that AIC penalization is less strong than BIC penalization (see Huang and Yang 2004), leading them to possess different properties. Whereas the BIC is consistent in the sense that if the true model is among the candidates the probability of selecting the true model approaches 1, the AIC possesses asymptotic optimality in the sense that if the true model is not among the candidates, the average squared error of the selected model will be asymptotically equivalent to the smallest possible offered by the candidate models. However, it seems to be well established (see Yang 2005) that if the true model is finite-dimensional, then the BIC should be preferred. Moreover, Huang and Yang (2004) proved that the BIC was able to select the correct lags within spline estimation under the constraints that the number of basis functions grew with the sample size. As the LHMs of the SNAKE model do not necessarily grow with the sample size but rather with the true underlying particularities of the data, in this paper, the behavior of the BIC will be compared against the 1-SER for model selection criterion within the SNAKE model.

Model selection via the BIC will be to select the model $S_s$ with the smallest BIC, defined as

$$\text{BIC}(S_s, \text{GS+PS}) = \log(\text{MSE}(S_s, \text{GS+PS})) + \frac{N_s}{n}\log(n), \tag{10}$$

where $N_s$ is the dimension of the space of functions given by Eq. (2) that can be calculated as $N_s = 1 + \sum_{i \in s}(1 + K_i)$ (see Huang and Yang 2004).

## 4 Interpreting the SNAKE model

After fitting the model to the data, we can examine the roles of each input in modeling the output. One salient feature of the SNAKE model is its interpretability, inherited from the simple structure of additive models, and exploited by means of a particular way of plotting the SNAKE model's components in order to facilitate their exploration and interpretation. Basically, each term of Eq. 2 is plotted in a different graph, showing the estimated one-dimensional piecewise linear model associated with the identified input.

Consider, for instance, the straightforward case of Fig. 3, where the SNAKE model for one of the simulation processes is represented (furtherly studied in Sect. 5). Concretely, the process follows the (unknown) linear equation $y_t = -0.5y_{t-6} + 0.5y_{t-10} + 0.1\xi_t$ where $\xi_t$ is an i.i.d. $N(0, 1)$

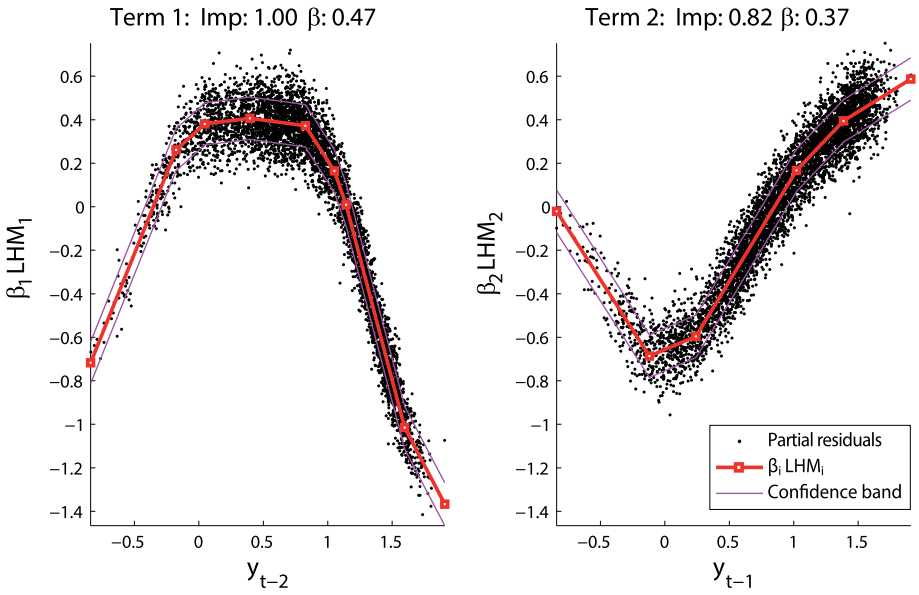


**Fig. 4** SNAKE model of a nonlinear process. Its mathematical expression is $y_t \approx 0.87 + 0.47\text{LHM}_1(y_{t-2}) + 0.37\text{LHM}_2(y_{t-1})$. These two nonlinear terms have been automatically estimated from the first 10 lags

random process. A quick review of the model representation allows obtaining simple conclusions about the problem. The output depends only on lags 6 and 10. Furthermore, there is a inverse linear relationship with lag 6, whereas it is direct with lag 10.

The general representation shown in the example of Fig. 3 has special features to be described. Note that this representation will be used in the rest of the paper to show the SNAKE models. Then, for each input variable a graph is plotted representing the term $\beta_i \text{LHM}_i$. This way, one can obtain the value of $\hat{y}_t$ by simply evaluating the straight lines (squares indicating the position of the knots) with the known values of the input variables plus adding $\beta_0$. In this case, $\hat{y}_t \approx 0.00 - 0.50y_{t-6} + 0.50y_{t-10}$. Also, normalized importances and each $\beta_i$ are represented above each term. The partial residuals are also plotted (black dots), which have been used to obtain each $\text{LHM}_i$. Finally, above and below the LHMs two piecewise linear functions are represented that give an idea of the distribution of the variance of the partial residuals around the LHM. Their shape actually follows LHM models of the conditional residuals of $\text{LHM}_i$, and their height is above and below one standard deviation from $\text{LHM}_i$.

Visualizing the model in this form allows to have a fast deep understanding of the physical behavior of the time series. This way, it is easy to observe if there exist any nonlinearities and in what an extend. For sufficiently weakly dependent lags, one can obtain simple mind rules about the increase or decrease in $y_t$ given variations on the significant lag values. Consider Fig. 4, where a SNAKE model from a nonlinear time series is represented. The particular underlying function is $y_t = -0.4(3 - y_{t-1}^2)/(1 + y_{t-1}^2) + 0.6[3 - (y_{t-2} - 0.5)^3]/[1 + (y_{t-2} - 0.5)^4] + 0.1\xi_t$ which in fact is not quite easy to interpret. However, after observing the SNAKE model, one can extract some useful information about the relationship between the output and the different lags. According to this model, $y_t$ has a nonlinear dependence on $y_{t-2}$ and $y_{t-1}$. When $y_{t-2}$ moves away from the interval $\approx [0, 0.75]$, then $y_t$ tends to decrease quickly. This effect is combined with the, less important, nonlinear response of $y_t$ to $y_{t-1}$. If $y_{t-1} < 0$, then the relationship between $y_t$ and $y_{t-1}$ is inverse, otherwise direct.

## 5 Simulation study

In order to evaluate the performance of the proposed approach, synthetic time series have been generated and the learning algorithm has been run to fit this data. In this way, the goodness of the variable selection system can be analyzed with perfect information. Learning and validation errors are presented for the different time series as well as a comparison with other well-known benchmark methodologies.

Hundred realizations of different series found in the literature have been simulated in a Monte Carlo procedure (see Table 1). The most well known of these time series, (AR1-AR3) and (NLAR1-NLAR3) were used by Tschernig and Yang (2000). NLAR4 was taken from Chen et al. (1995), while Huang and Yang (2004) introduced the lasting two (NLAR1U1 and NLAR1U2).

For sample sizes $n = 200,\ 500,\ 5000$, realizations of $n + 400$ were obtained and the first 400 observations were dismissed for ensuring stationarity. Results on Table 2 report the number of realizations that were underfitted (not selecting all the significant lags), correctly fitted (selecting all the significant lags and no other), or overfitted (selecting all the

**Table 1** Time series generating functions for the simulation study

| Problem | Function |
|---|---|
| AR1 | $y_t = 0.5y_{t-1} + 0.4y_{t-2} + 0.1\xi_t$ |
| AR2 | $y_t = -0.5y_{t-1} + 0.4y_{t-2} + 0.1\xi_t$ |
| AR3 | $y_t = -0.5y_{t-6} + 0.5y_{t-10} + 0.1\xi_t$ |
| NLAR1 | $y_t = -0.4(3 - y_{t-1}^2)/(1 + y_{t-1}^2) + 0.6[3-(y_{t-2}-0.5)^3]/[1+(y_{t-2}-0.5)^4]+0.1\xi_t$ |
| NLAR2 | $y_t = [-0.4 - 2\exp(-50y_{t-6}^2)]y_{t-6} + [0.5 - 0.5\exp(-50y_{t-10}^2)]y_{t-10} + 0.1\xi_t$ |
| NLAR3 | $y_t = y_{t-6}[-0.4 - 2\cos(40y_{t-6})\exp(-30y_{t-6}^2)] + y_{t-10}[0.55 - 0.55\sin(40y_{t-10})\exp(-10y_{t-10}^2)] + 0.1\xi_t$ |
| NLAR4 | $y_t = -2y_{t-1}I(y_{t-1} \leq 0) + 0.4y_{t-1}I(y_{t-1} > 0) + 0.1\xi_t$ |
| NLAR1U1 | $y_t = -0.4(3 - y_{t-1}^2)/(1 + y_{t-1}^2) + 0.1\xi_t$ |
| NLAR2U2 | $y_t = 0.6[3-(y_{t-2}-0.5)^3]/[1 + (y_{t-2}-0.5)^4] + 0.1\xi_t$ |

Note that $I(x)$ is an indicator which takes a value 1 if $x$ holds and 0 otherwise

significant lags but also others). For this experiment, the first 10 significant lags were considered as candidate variables.

It can be stated that the 1-SER criterion is stricter than the BIC in terms of allowing more terms in the model (none of the underfitting errors included any nonrelevant lag), and in general behaves poorly compared with the BIC, which provides fairly optimal results. Also, errors committed by the BIC are mainly biased upon overfitting, which discards the implementation of the AIC (which is known to overfit more easily). SNAKE models for one particular realization of AR3 and NLAR1 series are represented in Figs. 3 and 4, respectively, (see comments in Sect. 4). Hereafter are represented the remaining SNAKE models fitted for series from Table 1 ($n = 5000$).

Concerning the NLAR2 problem, although a first glance at the true underlying function of Table 1 suggests a similar complex structure for the significant lags, the SNAKE model of (Fig. 5) captures a clear nonlinear response of $y_t$ to $y_{t-6}$ and $y_{t-10}$, being more complex with lag 6.

For the NLAR3 time series, some special attention must be paid. First of all, notice from Table 2 that the BIC obtained the poorer results of all series. Although the model of Fig. 7 captures the main relationships between the output and its lags, there is remaining structure in the scatterplot of the partial residuals for the first term. An examination of the true underlying surface (Fig. 6) allows us to observe the intrinsic complexity. With the present noise and the small values of $n$, the LHM is not capable of detecting all the local minimums and maximums, which leads to an incorrect number of knots, which finally corrupts the estimation of both the MSE and $N_s$.

**Table 2** Lag selection results of the proposed approach, where for each criterion the first, second and third columns represent, respectively, the number of underfitted (UF), correctly fitted (CF), and overfitted (OF) realizations out of a total of 100

| Problem | $n$ | Results 1-SER | | | Results BIC | | |
|---|---|---|---|---|---|---|---|
| | | UF | CF | OF | UF | CF | OF |
| AR1 | 200 | 52 | 48 | 0 | 28 | 71 | 1 |
| | 500 | 20 | 80 | 0 | 0 | 100 | 0 |
| | 5000 | 25 | 75 | 0 | 0 | 100 | 0 |
| AR2 | 200 | 50 | 49 | 1 | 30 | 69 | 1 |
| | 500 | 20 | 80 | 0 | 0 | 100 | 0 |
| | 5000 | 21 | 79 | 0 | 0 | 100 | 0 |
| AR3 | 200 | 4 | 94 | 2 | 0 | 97 | 3 |
| | 500 | 0 | 100 | 0 | 0 | 100 | 0 |
| | 5000 | 1 | 99 | 0 | 0 | 100 | 0 |
| NLAR1 | 200 | 0 | 98 | 2 | 0 | 97 | 3 |
| | 500 | 0 | 100 | 0 | 0 | 98 | 2 |
| | 5000 | 0 | 100 | 0 | 0 | 100 | 0 |
| NLAR2 | 200 | 50 | 47 | 3 | 9 | 90 | 1 |
| | 500 | 17 | 83 | 0 | 0 | 100 | 0 |
| | 5000 | 14 | 86 | 0 | 0 | 98 | 2 |
| NLAR3 | 200 | 9 | 82 | 9 | 0 | 90 | 10 |
| | 500 | 0 | 98 | 2 | 0 | 94 | 6 |
| | 5000 | 0 | 98 | 2 | 0 | 93 | 7 |
| NLAR4 | 200 | 10 | 90 | 0 | 0 | 100 | 0 |
| | 500 | 17 | 82 | 1 | 0 | 100 | 0 |
| | 5000 | 20 | 79 | 1 | 0 | 100 | 0 |
| NLAR1U1 | 200 | 0 | 100 | 0 | 0 | 100 | 0 |
| | 500 | 0 | 100 | 0 | 0 | 100 | 0 |
| | 5000 | 0 | 100 | 0 | 0 | 100 | 0 |
| NLAR1U2 | 200 | 0 | 100 | 0 | 0 | 100 | 0 |
| | 500 | 0 | 100 | 0 | 0 | 100 | 0 |
| | 5000 | 0 | 100 | 0 | 0 | 99 | 1 |

Also, NLAR4 drives special attention as it appertains to the *self-exciting threshold autoregressive models* class (SETAR, see Tong 1993). According to the SNAKE model of Fig. 8, there is a different linear response depending on the sign of $y_{t-1}$, being the output response more sensitive to changes in $y_{t-1}$ when $y_{t-1} < 0$. Note that the structure of the SNAKE model is able to perfectly capture a threshold behavior, but only when the threshold variable coincides with the regression variable (as in the NLAR4 problem). If this is not the case, the interaction between variables should be included by other means.

It is interesting to observe SNAKE models for problems NLAR1U1 and NLAR1U2. Despite being both generated from a clearly nonlinear underlying function, the stochastic equilibrium of the function plus $0.1\xi_t$ leads the data to behave in very different ways. In the case of NLAR1U1,

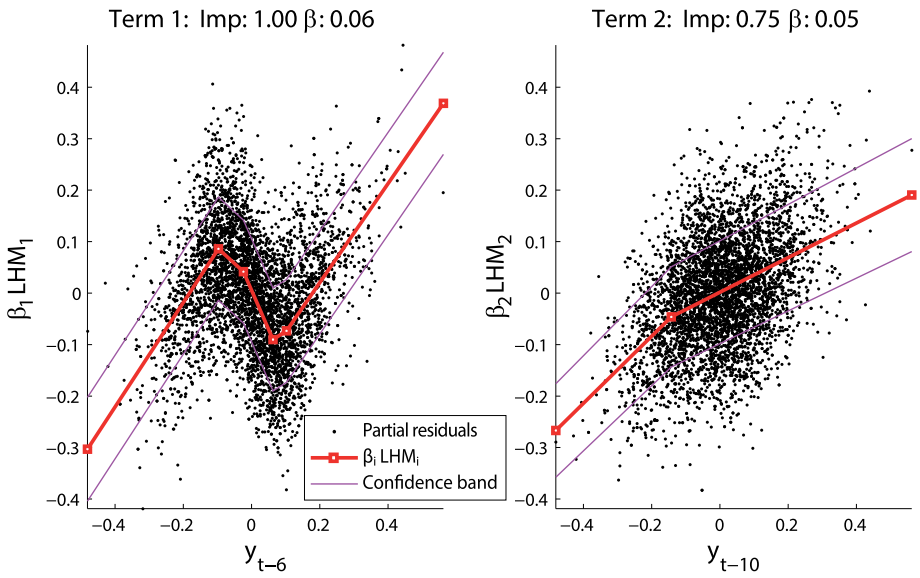


**Fig. 5** SNAKE model for the NLAR2 problem (see Table 1). This model consists of two terms, without constant, i.e., $y_t \approx 0.06\text{LHM}_1(y_{t-6}) + 0.05\text{LHM}_2(y_{t-10})$

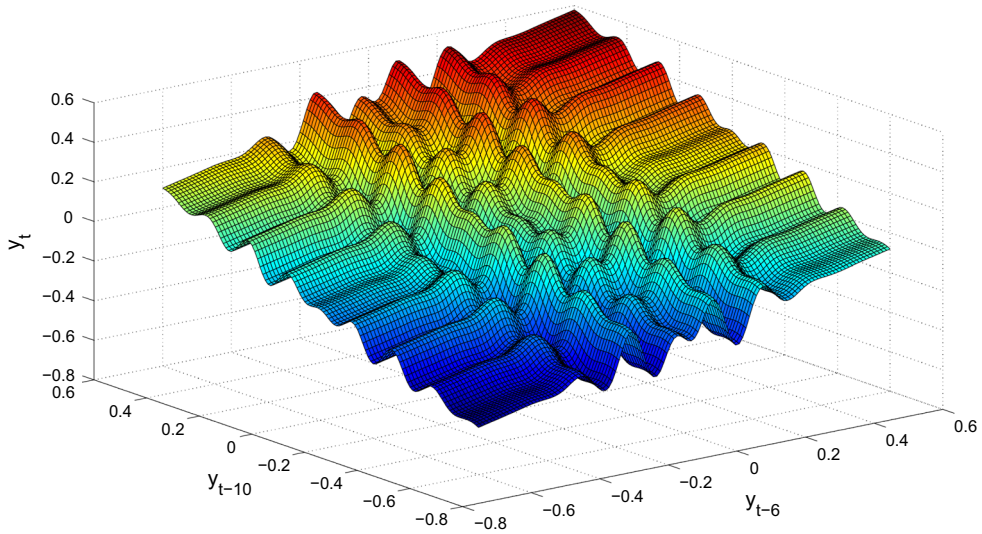


**Fig. 6** True underlying surface for the NLAR3 problem

as seen in Fig. 9, the time series is much like within an inverse linear relationship between $y_t$ and $y_{t-1}$, whereas for NLAR1U2 problem, the SNAKE model of Fig. 10 suggests a highly nonlinear relationship with two very different states. In particular, the scatterplot of the residuals suggests a two-state output response, where the output changes abruptly from state A to B every two instants. Note that the model also implies more uncertainty in $y_t$ when $y_{t-2}$ is in state B than in state A. It is the great interpretability of SNAKE's methodology that makes us be aware of these aspects, giving us the opportunity to analyze the logic governing the model and, if required, to choose a more appropriate model.

Finally, the SNAKE model has been compared with a linear AR model and a feed-forward neural network (multilayer perceptron, MLP) for all time series but NLAR1U1 and NLAR1U2, as they are mere simplifications of NLAR1. For each realization of the time series, the three kind of models were fitted. Table 3 shows the median errors made on the learning set (for $n$ being 200, 500 and 5000), and Table 4 shows the errors of these models on out-of-sample validation sets of 1000 length. RMSE is calculated as the square root of the MSE (Eq. 4), and we use the common definition of the mean absolute error (MAE) and $R^2$, given by

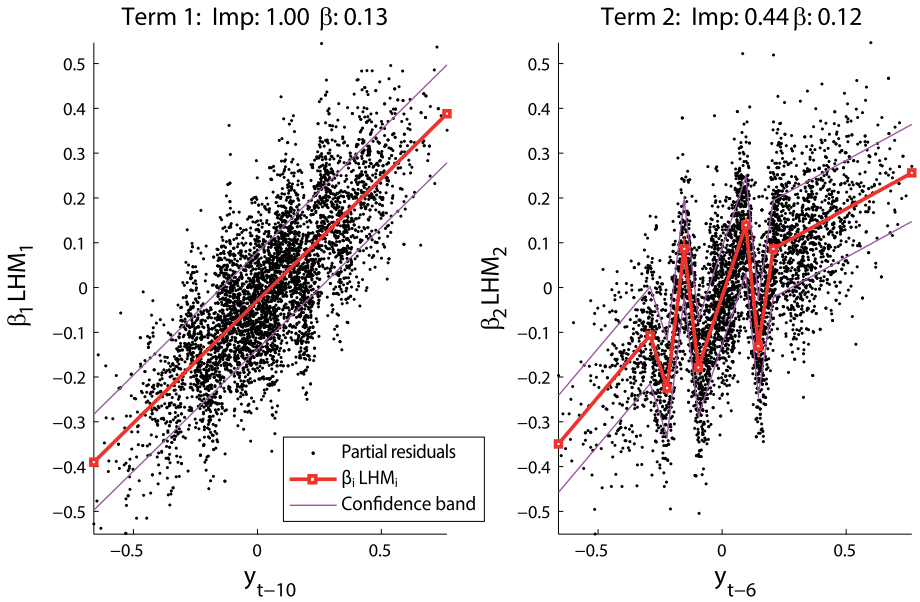


**Fig. 7** SNAKE model for the NLAR3 problem (see Table 1 and Fig. 6). This model has two main terms: $y_t \approx 0.05 + 0.13\text{LHM}_1(y_{t-10}) + 0.12\text{LHM}_2(y_{t-6})$

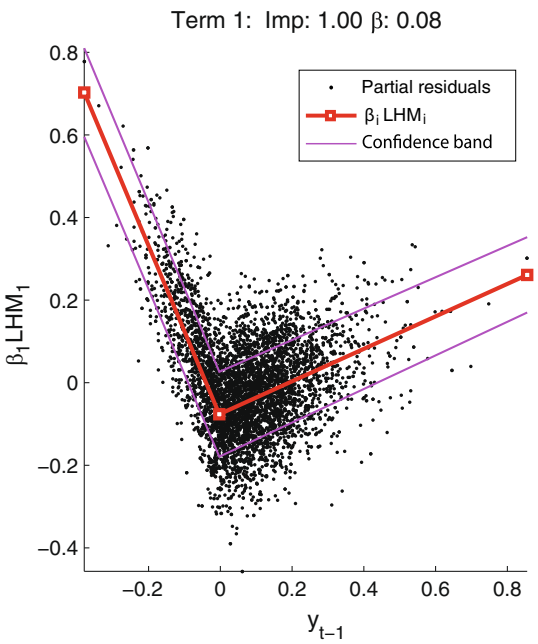


**Fig. 8** SNAKE model for the NLAR4 problem (see Table 1), given by $y_t \approx 0.08 + 0.08\text{LHM}_1(y_{t-1})$, where the nonlinear term has been automatically selected from the first 10 lags

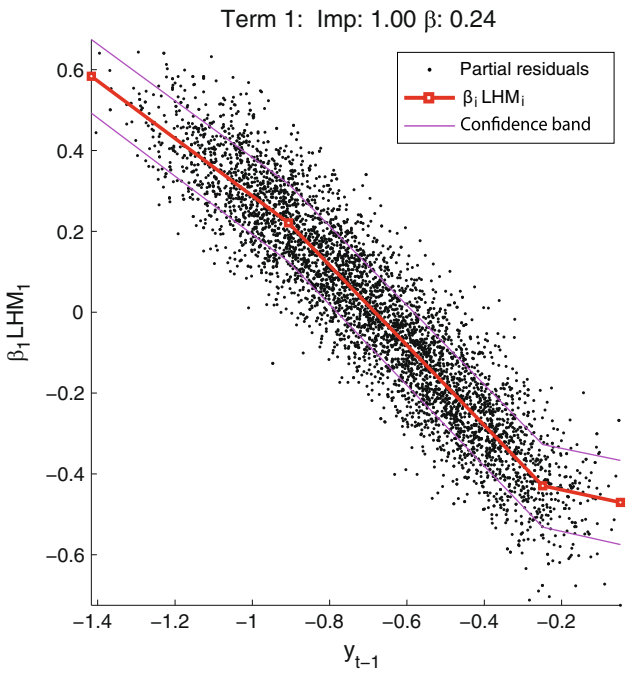


**Fig. 9** SNAKE model for the NLAR1U1 problem (see Table 1), automatically selected from the first 10 lags. This model can be expressed as $y_t \approx -0.69 + 0.24\text{LHM}_1(y_{t-1})$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - g(\boldsymbol{w_t})|\,, \tag{11}$$

$$\text{R}^2 = 1 - \frac{\text{MSE}}{\sigma^2(y_t)}, \tag{12}$$

where $g(\cdot)$ is the output estimated by the model, and $\sigma^2(y_t)$ is the sample variance of the time series. Note that, as the 10 first significant lags where considered as candidate variables, the actual size used for the training set is $n - 10$.

Identification of the significant lags was manually done for the AR and the MLP models. In particular, uncertainty due to lag selection has been completely removed in these two models by setting as input variables only the correct lags. This way, AR and MLP models have been biased toward non-overfitted but accurate benchmark models. Additionally, the MLP model was fully connected, with one hidden layer. The activation function of the neurons in the hidden layer was the hyperbolic tangent and the linear function for the output layer.

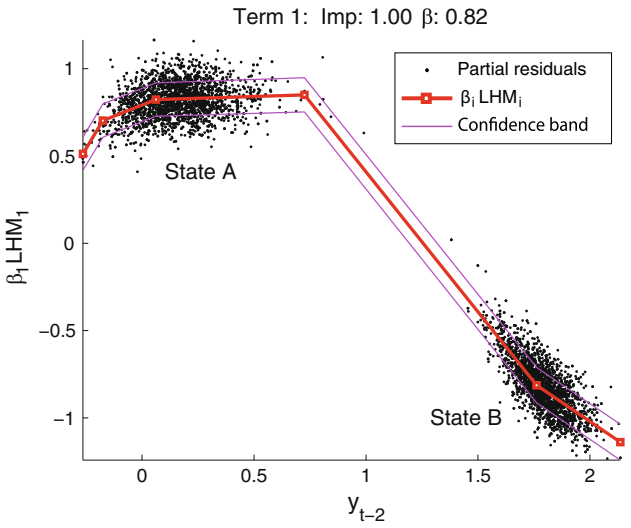


**Fig. 10** SNAKE model for the NLAR1U2 problem (see Table 1), automatically selected from the first 10 lags. This model can be expressed as $y_t \approx 0.96 + 0.82\text{LHM}_1(y_{t-2})$

One single neuron on the hidden layer was configured on the linear realizations (AR1 to AR3), while 5 neurons on the hidden layer were selected for NLAR1 and NLAR2. Because of the complexity of NLAR3 and the uncertainty given by the present noise, after user-analysis, 2, 5, and 20 hidden neurons were used for $n = 200, 500, 5000$, respectively, in order to avoid overfitting while capturing the highest information available. Two hidden neurons were used for NLAR4. Finally, note that the median errors reported about SNAKE models also include those that were incorrectly identified (using BIC).

From Table 3 and in comparison with Table 4, it can be stated that the models identified do not suffer from overfitting (neither from underfitting). The expected value of RMSE would be 0.1 in an ideally fitted model, due to noise $0.1\xi_t$. It can be concluded that the SNAKE model is capable of obtaining as good results as linear models when dealing with linear time series and as good results as well-known models like neural networks when dealing with nonlinear time series.

## 6 Forecasting Spanish residential natural gas consumption

In order to illustrate SNAKE's methodology in a real case study, we have applied it to one-day-ahead forecasting of the natural gas demand of a particular region of Spain, where most of the gas consumption comes from residential and

**Table 3** Simulation median results in learning set, obtained from 100 replications

| | $n$ | MAE | | | RMSE | | | $R^2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AR | SNAKE | MLP | AR | SNAKE | MLP | AR | SNAKE | MLP |
| AR1 | 200 | 0.0786 | 0.0786 | 0.0781 | 0.0988 | 0.0992 | 0.0983 | 0.6892 | 0.6722 | 0.6938 |
| | 500 | 0.0800 | 0.0798 | 0.0798 | 0.0999 | 0.0992 | 0.0998 | 0.7332 | 0.7345 | 0.7332 |
| | 5000 | 0.0797 | 0.0796 | 0.0797 | 0.0999 | 0.0996 | 0.0999 | 0.7375 | 0.7389 | 0.7376 |
| AR2 | 200 | 0.0788 | 0.0776 | 0.0783 | 0.0989 | 0.0988 | 0.0987 | 0.7377 | 0.6993 | 0.7427 |
| | 500 | 0.0792 | 0.0788 | 0.0791 | 0.0989 | 0.0986 | 0.0987 | 0.7257 | 0.7265 | 0.7276 |
| | 5000 | 0.0797 | 0.0791 | 0.0796 | 0.0999 | 0.0990 | 0.0999 | 0.7471 | 0.7474 | 0.7472 |
| AR3 | 200 | 0.0872 | 0.0779 | 0.0782 | 0.1091 | 0.0979 | 0.0981 | 0.6997 | 0.7499 | 0.7481 |
| | 500 | 0.0864 | 0.0787 | 0.0791 | 0.1077 | 0.0984 | 0.0985 | 0.6999 | 0.7536 | 0.7520 |
| | 5000 | 0.0874 | 0.0794 | 0.0797 | 0.1096 | 0.0994 | 0.0998 | 0.6981 | 0.7539 | 0.7505 |
| NLAR1 | 200 | 0.2849 | 0.0763 | 0.0765 | 0.3504 | 0.0964 | 0.0959 | 0.5313 | 0.9655 | 0.9656 |
| | 500 | 0.2910 | 0.0782 | 0.0783 | 0.3581 | 0.0980 | 0.0980 | 0.5282 | 0.9643 | 0.9639 |
| | 5000 | 0.2897 | 0.0783 | 0.0801 | 0.3572 | 0.0985 | 0.1005 | 0.5258 | 0.9641 | 0.9624 |
| NLAR2 | 200 | 0.0948 | 0.0784 | 0.0735 | 0.1188 | 0.0981 | 0.0929 | 0.1576 | 0.4094 | 0.4675 |
| | 500 | 0.0940 | 0.0793 | 0.0777 | 0.1170 | 0.0994 | 0.0972 | 0.1704 | 0.4014 | 0.4277 |
| | 5000 | 0.0951 | 0.0801 | 0.0798 | 0.1188 | 0.1000 | 0.1002 | 0.1599 | 0.4023 | 0.4065 |
| NLAR3 | 200 | 0.1065 | 0.0869 | 0.0991 | 0.1333 | 0.1100 | 0.1254 | 0.5556 | 0.7129 | 0.6000 |
| | 500 | 0.1088 | 0.0871 | 0.0981 | 0.1353 | 0.1094 | 0.1233 | 0.6203 | 0.7543 | 0.6875 |
| | 5000 | 0.1096 | 0.0871 | 0.0821 | 0.1368 | 0.1097 | 0.1030 | 0.6784 | 0.7536 | 0.8146 |
| NLAR4 | 200 | 0.0951 | 0.0783 | 0.0797 | 0.1210 | 0.0985 | 0.0999 | 0.0358 | 0.3688 | 0.3484 |
| | 500 | 0.0960 | 0.0776 | 0.0804 | 0.1215 | 0.0976 | 0.1008 | 0.0286 | 0.3552 | 0.3370 |
| | 5000 | 0.0974 | 0.0787 | 0.0806 | 0.1238 | 0.0984 | 0.1010 | 0.0262 | 0.3709 | 0.3537 |

**Table 4** Out-of-sample median simulation results obtained from 100 replications

| | $n$ | MAE | | | RMSE | | | $R^2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AR | SNAKE | MLP | AR | SNAKE | MLP | AR | SNAKE | MLP |
| AR1 | 200 | 0.0807 | 0.0862 | 0.0822 | 0.1008 | 0.1094 | 0.1036 | 0.7200 | 0.6726 | 0.7003 |
| | 500 | 0.0799 | 0.0807 | 0.0803 | 0.0999 | 0.1010 | 0.1005 | 0.7312 | 0.7219 | 0.7289 |
| | 5000 | 0.0797 | 0.0810 | 0.0797 | 0.0999 | 0.1017 | 0.0998 | 0.7284 | 0.7201 | 0.7285 |
| AR2 | 200 | 0.0802 | 0.0835 | 0.0810 | 0.1007 | 0.1060 | 0.1015 | 0.7388 | 0.6998 | 0.7349 |
| | 500 | 0.0804 | 0.0812 | 0.0808 | 0.1007 | 0.1020 | 0.1012 | 0.7429 | 0.7356 | 0.7412 |
| | 5000 | 0.0799 | 0.0809 | 0.0799 | 0.0998 | 0.1016 | 0.0999 | 0.7316 | 0.7229 | 0.7310 |
| AR3 | 200 | 0.0877 | 0.0829 | 0.0813 | 0.1102 | 0.1044 | 0.1033 | 0.6893 | 0.7152 | 0.7336 |
| | 500 | 0.0883 | 0.0815 | 0.0806 | 0.1107 | 0.1027 | 0.1023 | 0.6926 | 0.7246 | 0.7387 |
| | 5000 | 0.0880 | 0.0810 | 0.0803 | 0.1101 | 0.1018 | 0.1019 | 0.6933 | 0.7317 | 0.7408 |
| NLAR1 | 200 | 0.2911 | 0.0871 | 0.0888 | 0.3613 | 0.1104 | 0.1126 | 0.5167 | 0.9543 | 0.9528 |
| | 500 | 0.2918 | 0.0835 | 0.0832 | 0.3599 | 0.1057 | 0.1044 | 0.5147 | 0.9584 | 0.9596 |
| | 5000 | 0.2909 | 0.0831 | 0.0813 | 0.3567 | 0.1054 | 0.1018 | 0.5282 | 0.9589 | 0.9619 |
| NLAR2 | 200 | 0.0964 | 0.0838 | 0.0878 | 0.1202 | 0.1051 | 0.1109 | 0.1459 | 0.3405 | 0.2516 |
| | 500 | 0.0961 | 0.0822 | 0.0836 | 0.1197 | 0.1031 | 0.1052 | 0.1497 | 0.3642 | 0.3464 |
| | 5000 | 0.0952 | 0.0819 | 0.0803 | 0.1188 | 0.1031 | 0.1005 | 0.1506 | 0.3655 | 0.3937 |
| NLAR3 | 200 | 0.1128 | 0.1065 | 0.1092 | 0.1407 | 0.1342 | 0.1367 | 0.6204 | 0.6340 | 0.6178 |
| | 500 | 0.1113 | 0.0945 | 0.1071 | 0.1390 | 0.1191 | 0.1351 | 0.6418 | 0.7325 | 0.6581 |
| | 5000 | 0.1102 | 0.0956 | 0.0841 | 0.1373 | 0.1199 | 0.1054 | 0.6475 | 0.7280 | 0.7823 |
| NLAR4 | 200 | 0.0980 | 0.0811 | 0.0824 | 0.1250 | 0.1019 | 0.1029 | 0.0146 | 0.3397 | 0.3198 |
| | 500 | 0.0982 | 0.0818 | 0.0824 | 0.1248 | 0.1026 | 0.1032 | 0.0200 | 0.3391 | 0.3368 |
| | 5000 | 0.0974 | 0.0813 | 0.0809 | 0.1241 | 0.1021 | 0.1015 | 0.0249 | 0.3443 | 0.3500 |

commercial use. In Spain, residential and commercial natural gas consumption by end use is primarily linked, as usual, with heating (including hot water) and cooking.

With the aim of forecasting residential and commercial natural gas demand of a region, different exogenous variables can be considered depending on the forecasting horizon. In the particular case of short-term forecasting of daily natural gas demand, according to the review of Soldo (2012), the typical candidate exogenous variables can be grouped in two main types: weather-related and calendar-related variables. The first group consists of the wind speed and the outdoor temperature, being well known that temperature has the greatest impact on the consumption (see, e.g., Demirel et al. 2012). The second group is formed by exogenous variables derived from the calendar such as the day of the week, the month or the day of the year. It also includes indicator variables used to mark holidays or weekend days. Note that demographic and economic factors like the population, the number of customers, the price of gas or the Gross Domestic Product are, in general, not used in one-day-ahead forecasting of daily natural gas demand due to their very low dynamics (see, e.g., Soldo 2012).

In our particular case, regarding weather-related explanatory variables, we dismissed at the very beginning of the study the wind speed as a candidate exogenous variable because of its impact on consumption is negligible compared to temperature, (see, e.g., Szoplik 2015). Thus, focusing on temperature, two main exogenous variables have been considered: mean outdoor air temperature (calculated as $(T_{\max} + T_{\min})/2$ on a day) and the daily temperature range calculated as the difference between the maximum and the minimum temperature for each day. These two variables will be less correlated with each other than using $T_{\max}$ or $T_{\min}$, making it easier to use in statistical modeling. In this study, the objective is to make a short-term demand prediction of the total demand of the region for the next day. Taking this into account, lagged values of the demand and of the temperatures may also be important inputs. Also, the current week of the year has been found to be an important variable, good compromise between the current day of the year (too specific information) and the month of the year (too vague).

Fig. 11 shows the time series of the log-normalized demand, the mean temperature at the region, and the difference between the maximum and the minimum, including the expected mean temperature given by a reference model like the one on Sánchez-Úbeda and Berzosa (2011). The original demand obtained from Enagás has been normalized between

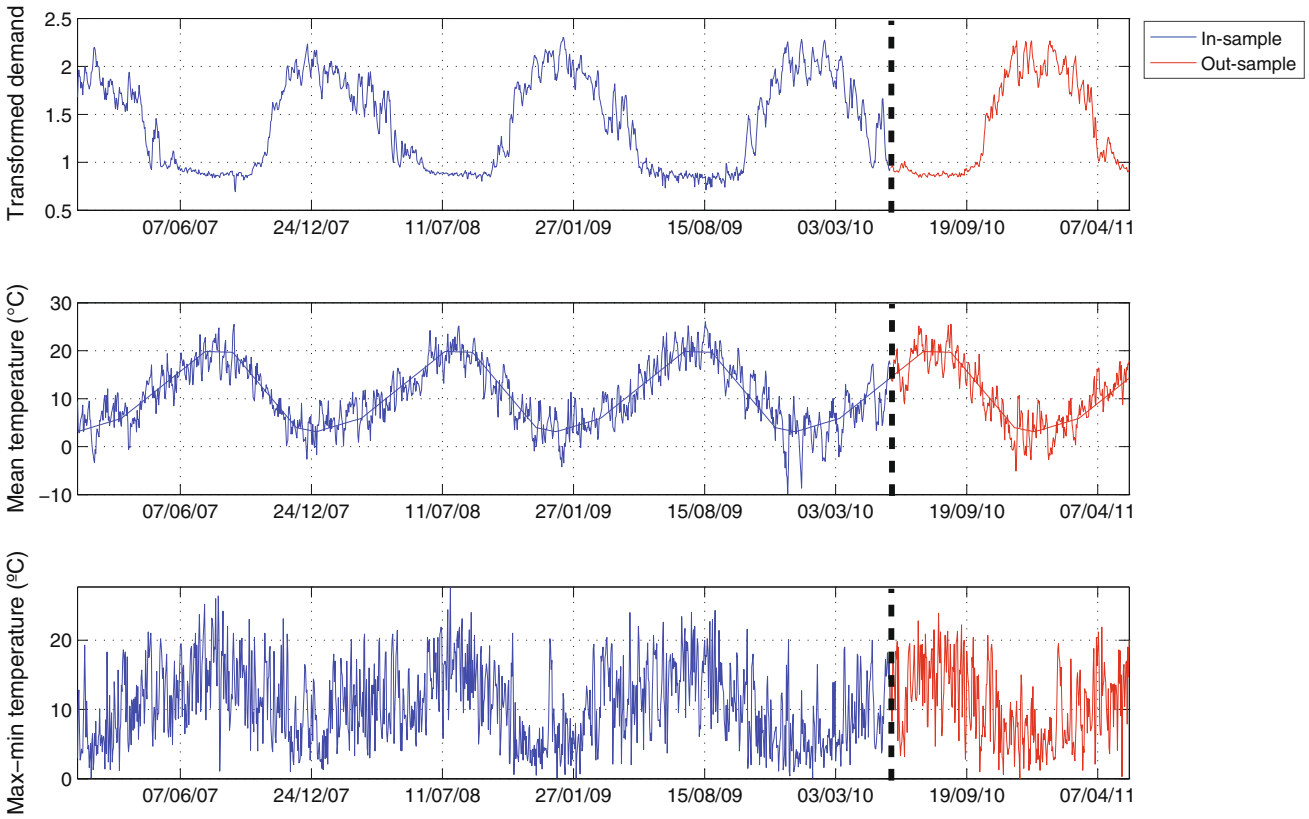


**Fig. 11** *Top* Log-normalized daily natural gas consumption at a particular region of Spain. *Middle* Associated mean outdoor air temperature, with mean expected temperature. *Bottom* Corresponding temperature range. In-sample and out-sample periods are shown in *blue* and *red*, respectively. (Color figure online)

10 and 2, and then a natural logarithm transformation has been applied to stabilize the variance. Data ranges from January 1, 2007 to May 24, 2011. To measure the quality of the forecasts, last 365 days have been reserved for validation. One clear feature of this demand is the strong seasonality throughout the year, mainly governed by the slow variation of the temperature (seasons). Other noticeable feature is that there is no trend over time.

On the other hand, fast variations of the temperature also have a great effect in the demand, as shown in Fig. 12 where it can be observed that the consumption is also inversely influenced by the rapid fluctuations of the temperature around the reference temperature. This effect happens only during a particular period of the year, when temperatures are lower (from October to May). Furthermore, no marked differences exist between working days and weekend demand.

To sum up, and considering all the aforementioned information, the SNAKE model has been adjusted for predicting one-step-ahead natural gas consumption. Three exogenous candidate variables were considered: the current week of the year, the mean temperature of the day ($T^{\text{mean}}$) and the daily temperature range ($T^{\text{rng}}$). Additionally, lags 1, 2, 3, 7, 8 and 9 of both the exogenous variables and the gas demand were considered as possibly important.

With all these 27 candidate variables, the SNAKE model was adjusted in the learning period, using the BIC as variable selection method and with parameters $\eta = 0.1$ and $\epsilon = 0$ for the LHM. The model obtained is shown in Fig. 13.

A first glance at the terms of the model provides us with several pieces of information. First, significant lags are 1 and 2 for the gas demand, and lag 1 for $T^{\text{mean}}$. This implies that there is not a relevant weekly seasonality. Also, the demand from the previous day seems to be the most important input, followed by the mean temperature. The week of the year seems to be also important, whereas $T^{\text{rng}}$ has been discarded. There exists a high correlation between two consecutive demand values, but is worth noticing that this correlation is not purely linear; in Fig. 14, the $y_{t-1}$ term is represented within an histogram of the partial residuals. A not spurious first piece is present, with a much lower slope and a high number of data. This piece represents low demands (typically associated with hot temperatures), where the existing relationship is much less marked. In particular, the bimodal histogram of the partial residuals shows that dur-

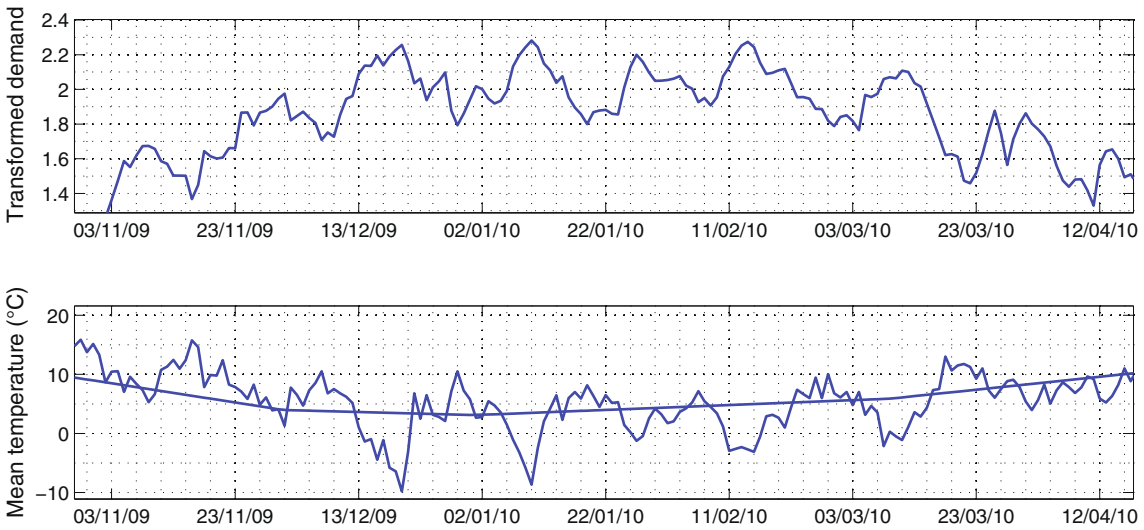


**Fig. 12** *Top* Log-normalized daily natural gas consumption detail in winter, at a particular region of Spain. *Bottom* Associated mean outdoor air temperature (with mean expected temperature)

ing high demands (associated with the low temperatures of winter), there is a clear linear relationship between $y_t$ and $y_{t-1}$. However, during non-cold seasons, with low demands, the relationship between $y_t$ and $y_{t-1}$ is much less marked, avoiding large changes in $y_t$ when $y_{t-1} < 0.9$.

On the other hand, according to the second term of the SNAKE model, there is the typical highly nonlinear response of natural gas demand to temperature (see, e.g., Szoplik 2015), where the existing inverse relationship between $y_t$ and $T_t^{\text{mean}}$ vanishes when $T_t^{\text{mean}}$ is larger than a particular threshold. In particular, for high temperatures (over 14 °C), natural gas house heating is normally turned off and the demand is uncorrelated with the temperature. Once that the temperature is below this "comfort zone," heating begins to grow inversely with temperature. Finally, for very low temperatures, there exists some saturation effect where no more heating power is available due to the limited capacity of the installed heating appliances, and changes in temperature do not reflect so important changes in demand. Note that the threshold value estimated by the model (14 °C) is quite similar to the standard 15 °C value used in Spain as base temperature for computing the HDD, but it has not been fixed a priori.

As usual, the effects of consecutive lag inputs are inversely proportional to the previous lags but with lesser importance, and this is reflected in terms $T_{t-1}^{\text{mean}}$ and $y_{t-2}$. Finally, the effect of the week in the year (calendar effect) can also be appreciated. This term reflects the higher demand on winter weeks, specially the effects that are not related to temperature but more to sociological aspects (e.g., in Spain, central house heating is usually turned on in mid-November for an stipulated number of hours, regardless of the outside temperature).

In order to be completely confident with SNAKE's automatic procedure and accuracy, a dynamic regression model (DR) and a neural network (MLP) were adjusted similarly to Sect. 5. The linear transfer function (LTF) methodology by Pankratz (1991) was used to identify an appropriate dynamic regression model, resulting in the model given by Eq. 13, where the $p$ values of all parameters are approximately zero. In order to reveal the underlying regression structure, the model has been expressed in its polynomial form, using the well-known lag (or backshift) operator $L^k y_t = y_{t-k}$. Note that in this case, $T^{\text{rng}}$ has found to be slightly significant, whereas the week of the year is not an adequate variable for this type of regression. Note also that there exist no moving average terms, which implies that SNAKE's NAARX approach will not be far from optimal.

$$y_t = \frac{-0.016}{1-0.57L^1} T_t^{\text{mean}} - 0.001 T_t^{\text{rng}} + \frac{1}{(1-L^1)(1+0.117L^2)(1+0.157L^3)} \varepsilon_t \tag{13}$$

Finally, an MLP model was adjusted that shares the same configuration as the ones on Sect. 5. The hidden layer possesses 25 neurons, and the selection of the most important inputs has been done via statistically sensitivity analysis. From the 1241 in-sample data observations, 993 random observations were divided into training and the other 248 were used for testing (used for cross-validation with early stopping). The process was repeated with different configurations (i.e., different number of neurons in the hidden layer) as well as different initializations for the same configuration (in order to avoid local minima). After a large manual

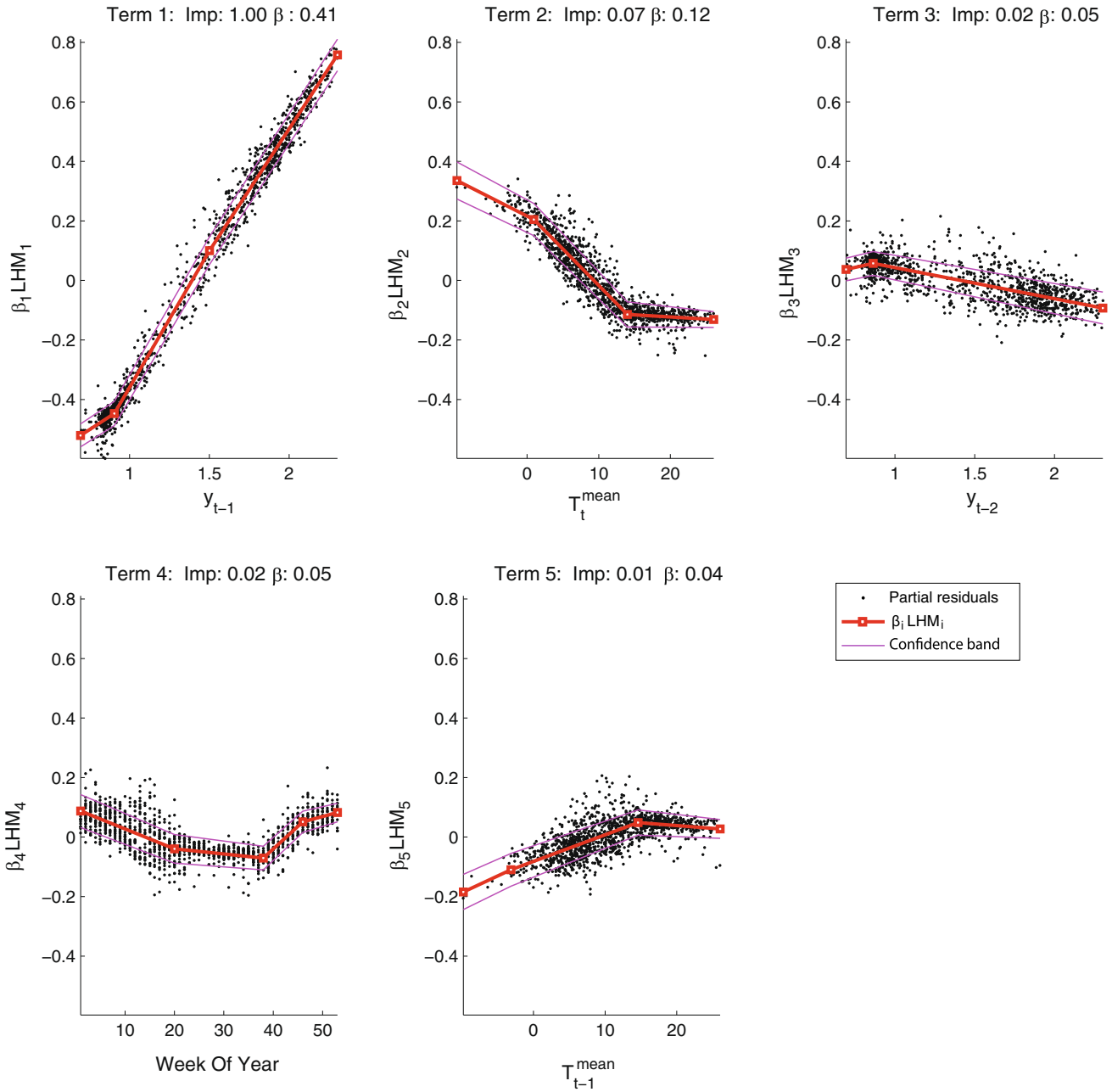


**Fig. 13** SNAKE model for the natural gas consumption problem. This model has five terms: $y_t \approx 1.40 + 0.41\text{LHM}_1(y_{t-1}) + 0.12\text{LHM}_2(T_t^{\text{mean}}) + 0.05\text{LHM}_3(y_{t-2}) + 0.05\text{LHM}_4(\text{WeekOfYear}) + 0.04\text{LHM}_5(T_{t-1}^{\text{mean}})$

process of identification including trial-and-error steps, the significant variables beginning with more than 20 candidate inputs have found to be the same that the ones present in the SNAKE model. This particular subset selection joins a low error in the learning set with a good generalization performance. Figure 15 shows 95% percentiles of the MLP's sensitivity absolute values normalized between 0 and 1 (see Muñoz and Czernichow 1998). Note the great similarity with the importances of the SNAKE model, provided in Fig. 13.

Final out-of-sample results are provided in Table 5, showing both the MAE and the RMSE for the three models. According to these figures, the SNAKE model is able to produce as good one-day-ahead forecasts of the natural gas demand as the MLP, while providing useful information about the possible nonlinear relationships between the output and the inputs.

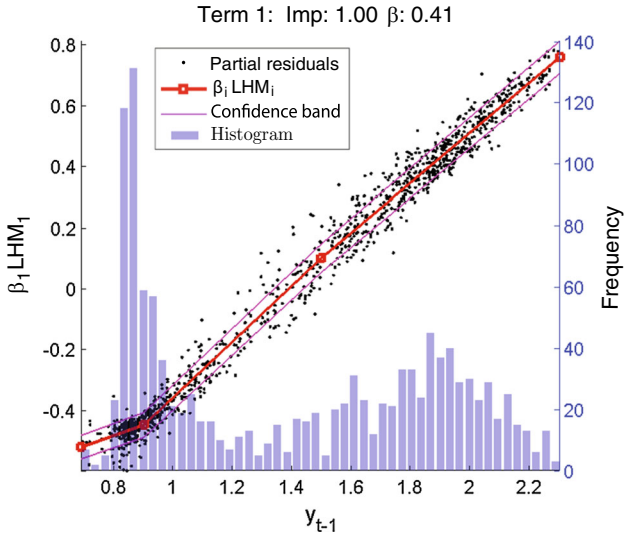


**Fig. 14** Detail of the first term of SNAKE model shown in Fig. 13, including the histogram of the partial residuals

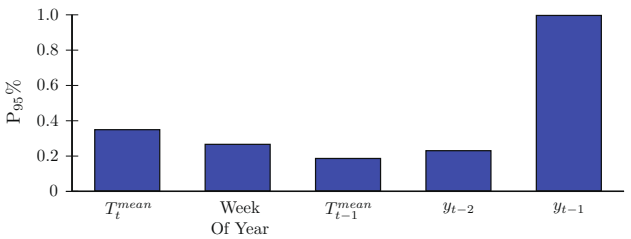


**Fig. 15** Normalized 95 percentiles of the input variables sensitivities

**Table 5** Out-of-sample prediction errors for the short-term natural gas demand

| | DR | SNAKE | MLP |
|---|---|---|---|
| MAE | 0.03654 | 0.02827 | 0.02827 |
| RMSE | 0.04687 | 0.03909 | 0.03970 |

## 7 Conclusions

In this article, an interpretable model has been proposed for nonlinear forecasting. Using a multivariate additive approximation with piecewise linear regression, it has been possible to present a methodology that will help the analyst in the labor of obtaining useful information from time series forecasting, as an straightforward physical interpretation may be attached to the components of the model. Moreover, the simulation study carried out using synthetic problems has proven that SNAKE models identified do not suffer from overfitting (neither from underfitting), being capable of obtaining as good results as linear models when dealing with linear time series and as good results as well-known models like neural networks when dealing with nonlinear time series.

The flexibility of the model has been widen compared to previous state-of-the-art approaches, as the number of knots in linear regression is not assumed given a sample size. Furthermore, the automatic model identification has proven to behave optimally through Monte Carlo simulations.

In the context of short-term natural gas demand forecasting, we have applied the proposed approach to one-day-ahead forecasting of the natural gas demand of a particular region of Spain. The SNAKE model is able to automatically perform as accurate predictions as the MLP model for this nonlinear problem, while relieving the researcher to make a manual identification process and allowing for a high understanding of the underlying physics of the problem.

Some area of possible future improvement over the SNAKE model would be to address the issue of correctly identifying moving average terms. Also, some kind of correction could be implemented when the additivity condition does not hold. Last but not least, facing all kind of problems, even where the generating processes are not stationary neither strongly mixing shall be a remaining goal for the automatic SNAKE model.

**Acknowledgements** The authors would like to acknowledge Enagás (Technical Manager of the Spanish gas system) for providing data for the case study.

## References

Akaike, H.: Statistical predictor identification. Ann. Inst. Stat. Math. **22**(1), 203–217 (1970). doi:10.1007/BF02506337, URL http://www.springerlink.com/content/ch4g3t5064842221/abstract/

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees, 1st edn. Chapman and Hall, New York (1984)

Chen, R., Tsay, R.S.: Nonlinear additive ARX models. J. Am. Stat. Assoc. **88**(423), 955–967 (1993). doi:10.2307/2290787, URL http://www.jstor.org/stable/2290787

Chen, R., Liu, J.S., Tsay, R.S.: Additivity test for nonlinear autoregression. Biometrika **80**, 369–383 (1995)

Cheng, B., Tong, H.: On consistent nonparametric order determination and chaos (with discussion). J. R. Stat. Soc. Ser. B Methodol **54**, 427–474 (1992)

Clements, M.P., Franses, P.H., Swanson, N.R.: Forecasting economic and financial time-series with non-linear models. Int. J. Forecast. **20**(2), 169–183 (2004). doi:10.1016/j.ijforecast.2003.10.004

Cruz, A., Muñoz, A., Luis Zamora, J., Espínola, R.: The effect of wind generation and weekday on spanish electricity spot price forecasting. Electric Power Syst. Res. **81**(10), 1924–1935 (2011). doi:10.1016/j.epsr.2011.06.002

Demirel, O.F., Zaim, S., Alikan, A., Zuyar, P.: Forecasting natural gas consumption in Istanbul using neural networks and multivariate time series methods. Turk. J. Electr. Eng. Comput. Sci. **20**(5), 695–711 (2012). www.scopus.com

Fan, S., Hyndman, R.J.: Short-term load forecasting based on a semi-parametric additive model. IEEE Trans. Power Syst. **27**(1), 134–141 (2012). doi:10.1109/TPWRS.2011.2162082

Friedman, J.H.: A variable span smoother. Tech. rep. (1984) URL http://stinet.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA148241

Gascón, A., Sánchez-Úbeda, E.F.: Application of multi-objective genetic algorithms to fitting piecewise linear models. In: 14th Conference of the Spanish Association for Artificial Intelligence (CAEPIA), Tenerife, Spain (2011)

Guo, Z., Shintani, M.: Nonparametric lag selection for nonlinear additive autoregressive models. Econ. Lett. **111**(2), 131–134 (2011). doi:10.1016/j.econlet.2011.01.014

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)

Huang, J.Z., Yang, L.: Identification of non-linear additive autoregressive models. J. R. Stat. Soc. Ser. B Stat. Methodol. **66**(2), 463–477 (2004)

Lewis, P.A.W., Stevens, J.G.: Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). J. Am. Stat. Assoc. **86**(416), 864–877 (1991). doi:10.2307/2290499, URL http://www.jstor.org/stable/2290499

Muñoz, A., Czernichow, T.: Variable selection using feedforward and recurrent neural networks. Eng. Intell. Syst. Electr. Eng. Commun. **6**(2), 91–102 (1998)

Muñoz, A., Sánchez-Úbeda, E.F., Cruz, A., Marín, J.: Short-term forecasting in power systems: a guided tour. In: Handbook of Power Systems II. Springer, pp. 129–160 (2010)

Pankratz, A.: Forecasting with Dynamic Regression Models. Wiley Series in Probability and Mathematical Statistics, New York, book, Whole (1991)

Sánchez-Úbeda, E.F.: Models for Data Analysis: Contributions to Automatic Learning. PhD thesis, Universidad Pontificia Comillas de Madrid, Madrid, (1999)

Sánchez-Úbeda, E.F., Berzosa, A.: Modeling and forecasting industrial end-use natural gas consumption. Energy Econ. **29**(4), 710–742 (2007). doi:10.1016/j.eneco.2007.01.015

Sánchez-Úbeda, E.F., Berzosa, A.: New variables to improve electricity and natural gas consumption forecasting: dynamic degree-days. In: 14th Conference of the Spanish Association for Artificial Intelligence (CAEPIA), Tenerife, Spain (2011)

Sánchez-Úbeda, E.F., Wehenkel, L.: The hinges model: a one-dimensional continuous piecewise polynomial model. In: Proceedings of the International Congress on Information Processing and Management of Uncertainty in Knowledge based Systems, IPMU98, Paris (1998)

Sánchez-Úbeda, E.F., Wehenkel, L.A.: Automatic fuzzy-rules induction by using the ORTHO model. In: Proceedings of Information Processing and Management of Uncertainty in Knowledge-based Systems, Madrid, pp. 1652–1659 (2000)

Soldo, B.: Forecasting natural gas consumption. Appl. Energy **92**, 26–37 (2012). doi:10.1016/j.apenergy.2011.11.003

Szoplik, J.: Forecasting of natural gas consumption with artificial neural networks. Energy **85**, 208–220 (2015). doi:10.1016/j.energy.2015.03.084, URL http://linkinghub.elsevier.com/retrieve/pii/S036054421500393X

Tjøstheim, D., Auestad, B.H.: Nonparametric identification of nonlinear time series: selecting significant lags. J. Am. Stat. Assoc. **89**(428), 1410–1419 (1994). doi:10.2307/2291003, URL http://www.jstor.org/stable/2291003

Tong, H.: Non-Linear Time Series: A Dynamical System Approach. Oxford University Press, Oxford (1993)

Tschernig, R., Yang, L.: Nonparametric lag selection for time series. J. Time Ser. Anal. **21**(4), 457–487 (2000). doi:10.1111/1467-9892.00193, URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9892.00193/abstract

Vieu, P.: Order choice in nonlinear autoregressive models. Statistics **26**(4), 307–328 (1995). doi:10.1080/02331889508802499, URL http://www.tandfonline.com/doi/abs/10.1080/02331889508802499#preview

Wehenkel, L.A.: Automatic Learning Techniques in Power Systems. Springer, Berlin (1998)

Yang, Y.H.: Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika **92**(4), 937–950 (2005). doi:10.1093/biomet/92.4.937

Yao, Q., Tong, H.: On subset selection in non-parametric stochastic regression. Stat. Sin. **4**(1), 51–70 (1994)

Zhu, L., Li, M.S., Wu, Q.H., Jiang, L.: Short-term natural gas demand prediction based on support vector regression with false neighbours filtered. Energy **80**, 428–436 (2015). doi:10.1016/j.energy.2014.11.083, URL http://www.sciencedirect.com/science/article/pii/S0360544214013553